Sun StorageTek 2540 / ZFS Performance Summary

Bob Friesenhahn
Simple Systems
bfriesen@GraphicsMagick.org
February 26, 2008

This white-paper is a summary of my activities related to setting up, tuning, and performance testing a Sun StorageTek 2540 drive array containing 12 300GB 15K RPM SAS drives (XTA2540R01A2F3600) attached to a Sun Ultra-40M2 workstation with an Emulex LPe11002-M4 dual-port FC controller and 20GB of RAM. The objective is to achieve the best possible single-user sequential read/write performance without sacrificing data integrity. There is also the objective to utilize the Solaris 10 ZFS filesystem to the maximum extent possible in order to obtain the best possible performance, minimize the risk of data loss, and make administration easier.

The StorageTek 2540 is used in a way that its inventors did not really intend. Each disk is exported as a LUN to create a JBOD array but with a battery-backed write cache for each disk. ZFS is used to create a pool of six mirrors in order to obtain excellent performance and reliability.

The final performance results are substantially better than Sun's own benchmark for this product. The official benchmark claims large file single-stream write performance of 105MB/second. The initial write performance observed in my testing was 150MB/second. The performance of the final configuration described here achieves a write performance of 267MB/second.

I would like to thank Mertol Ozyoney <Mertol.Ozyoney@Sun.COM> and Joel Miller <Joel.Miller@Sun.COM> for their valuable assistance with tuning the StorageTek 2540 for maximum performance.

StorageTek 2540 Configuration Details

Created storage profile

> Name: RAW_SAS
>
> Description: One SAS disk
>
> RAID Level: RAID 0
>
> Segment Size: 128KB
>
> No. of Disks: 1
>
> Disk Type: SAS

Created 12 storage pools (Disk-01 to Disk-12), each of which contains a single disk, and consuming all the space on that disk.

Created 12 volumes (Disk-01 to Disk-12), each one taking the full space of the similarly named storage pool. Solaris automatically creates the raw devices as each LUN is created.

For each volume, the option "Write Cache With Replication Enabled" is changed from "True" to "False". This allows potential data loss if the power fails and the cache battery on one of the controllers fails, or a controller fails while it still contains cached data. Under normal operating conditions, each controller "owns" six drives. We will try to allocate our LUN pairs so that each drive is owned by a different controller and the default FC path is to the owning controller. If there is cache data loss

(resulting in data not being written to a drive), ZFS should manage it reasonably well (possibly without any loss at all) since we are exporting each disk as a LUN.

Disable NV cache sync in StorageTek 2540 drive array

The StorageTek 2540 foolishly flushes its battery-backed cache whenever it is requested to sync its cache. The CAM administrative interface does not provide a way to disable this behavior so a low-level firmware tweak is required as described below.

Date: Sat, 16 Feb 2008 08:09:11 PST

From: Joel Miller <joel.miller@sun.com>

To: zfs-discuss@opensolaris.org

Subject: Re: [zfs-discuss] Performance with Sun StorageTek 2540

Bob,

Here is how you can tell the array to ignore cache sync commands and the force unit access bits... (Sorry if it wraps..)

On a Solaris CAM install, the 'service' command is in "/opt/SUNWsefms/bin"

To read the current settings:

```
service -d arrayname -c read -q nvsram region=0xf2 host=0x00
```

save this output so you can reverse the changes below easily if needed...

To set new values:

```
service -d arrayname -c set -q nvsram region=0xf2 offset=0x17 value=0x01 host=0x00
service -d arrayname -c set -q nvsram region=0xf2 offset=0x18 value=0x01 host=0x00
service -d arrayname -c set -q nvsram region=0xf2 offset=0x21 value=0x01 host=0x00
```

Host region 00 is Solaris (w/Traffic Manager)

You will need to reboot both controllers after making the change before it becomes active.

-Joel

Enable multipathing for FC ports only

```
# stmsboot -e -D fp
```

List mult-path support interfaces

```
# mpathadm list mpath-support
mpath-support:  libmpscsi_vhci.so
```

List multi-path support properties (abbreviated)

```
# mpathadm show mpath-support libmpscsi_vhci.so
mpath-support:  libmpscsi_vhci.so
        Vendor:  Sun Microsystems
        Driver Name:  scsi_vhci
        Default Load Balance:  round-robin
        Supported Load Balance Types:
                round-robin
                logical-block
        Allows To Activate Target Port Group Access:  yes
        Allows Path Override:  no
```

```
        Supported Auto Failback Config:  1
        Auto Failback:  on
        Failback Polling Rate (current/max):  0/0
        Supported Auto Probing Config:  0
        Auto Probing:  NA
        Probing Polling Rate (current/max):  NA/NA
        Supported Devices:
                Vendor:  SUN
                Product:  Universal Xport
                Revision:
                Supported Load Balance Types:
                        round-robin

                Vendor:  SUN
                Product:  LCSM100_F
                Revision:
                Supported Load Balance Types:
                        round-robin
```

List initiator ports

```
# mpathadm list initiator-port
Initiator Port:  iqn.1986-03.com.sun:01:ba8803f0ffff.45ddd88b,4000002a00ff
Initiator Port:  10000000c967c82f
Initiator Port:  10000000c967c830
```

Check status of initiator port and remote port

Note that the StorageTek 2540 error counts jump dramatically every time the attached system is rebooted and are stable thereafter so the apparent high error count does not impact normal operation.

```
# fcinfo hba-port -l 10000000c967c830
HBA Port WWN: 10000000c967c830
        OS Device Name: /dev/cfg/c2
        Manufacturer: Emulex
        Model: LPe11002-M4
        Firmware Version: 2.72a2
        FCode/BIOS Version: none
        Type: L-port
        State: online
        Supported Speeds: 1Gb 2Gb 4Gb
        Current Speed: 4Gb
        Node WWN: 20000000c967c830
        Link Error Statistics:
                Link Failure Count: 0
                Loss of Sync Count: 2
                Loss of Signal Count: 0
                Primitive Seq Protocol Error Count: 0
                Invalid Tx Word Count: 0
                Invalid CRC Count: 0
# fcinfo remote-port -l -p 10000000c967c830
Remote Port WWN: 200400a0b83a8a0c
        Active FC4 Types:
        SCSI Target: yes
        Node WWN: 200400a0b83a8a0b
        Link Error Statistics:
                Link Failure Count: 6
                Loss of Sync Count: 22443
                Loss of Signal Count: 316
                Primitive Seq Protocol Error Count: 0
```

```
                Invalid Tx Word Count: 0
                Invalid CRC Count: 0
# fcinfo hba-port -l 10000000c967c82f
HBA Port WWN: 10000000c967c82f
        OS Device Name: /dev/cfg/c3
        Manufacturer: Emulex
        Model: LPe11002-M4
        Firmware Version: 2.72a2
        FCode/BIOS Version: none
        Type: L-port
        State: online
        Supported Speeds: 1Gb 2Gb 4Gb
        Current Speed: 4Gb
        Node WWN: 20000000c967c82f
        Link Error Statistics:
                Link Failure Count: 0
                Loss of Sync Count: 2
                Loss of Signal Count: 0
                Primitive Seq Protocol Error Count: 0
                Invalid Tx Word Count: 0
                Invalid CRC Count: 0
# fcinfo remote-port -l -p 10000000c967c82f
Remote Port WWN: 200500a0b83a8a0c
        Active FC4 Types:
        SCSI Target: yes
        Node WWN: 200400a0b83a8a0b
        Link Error Statistics:
                Link Failure Count: 5
                Loss of Sync Count: 23067
                Loss of Signal Count: 225
                Primitive Seq Protocol Error Count: 0
                Invalid Tx Word Count: 0
                Invalid CRC Count: 0
```

Obtain properties of each initiator port

```
# mpathadm show initiator-port 10000000c967c82f
Initiator Port:  10000000c967c82f
        Transport Type:  Fibre Channel
        OS Device File:  /devices/pci@79,0/pci10de,376@a/pci10df,fe00@0,1/fp@0,0
# mpathadm show initiator-port 10000000c967c830
Initiator Port:  10000000c967c830
        Transport Type:  Fibre Channel
        OS Device File:  /devices/pci@79,0/pci10de,376@a/pci10df,fe00@0/fp@0,0
```

List of multi-pathed logical units

```
# mpathadm list lu
        /scsi_vhci/disk@g600a0b800039c9b50000000000000000
                Total Path Count: 1
                Operational Path Count: 1
        /scsi_vhci/disk@g600a0b80003a8a0b0000000000000000
                Total Path Count: 1
                Operational Path Count: 1
        /dev/rdsk/c4t600A0B80003A8A0B0000096147B451BEd0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B800039C9B500000A9C47B4522Dd0s2
                Total Path Count: 2
                Operational Path Count: 2
```

```
        /dev/rdsk/c4t600A0B800039C9B500000AA047B4529Bd0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B80003A8A0B0000096647B453CEd0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B800039C9B500000AA447B4544Fd0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B80003A8A0B0000096A47B4559Ed0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B800039C9B500000AA847B45605d0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B80003A8A0B0000096E47B456DAd0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B800039C9B500000AAC47B45739d0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B800039C9B500000AB047B457ADd0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B80003A8A0B0000097347B457D4d0s2
                Total Path Count: 2
                Operational Path Count: 2
        /dev/rdsk/c4t600A0B800039C9B500000AB447B4595Fd0s2
                Total Path Count: 2
                Operational Path Count: 2
```

Details for a single logical unit

```
# mpathadm show lu /dev/rdsk/c4t600A0B800039C9B500000AB047B457ADd0s2
Logical Unit:  /dev/rdsk/c4t600A0B800039C9B500000AB047B457ADd0s2
        mpath-support:  libmpscsi_vhci.so
        Vendor:  SUN
        Product:  LCSM100_F
        Revision:  0617
        Name Type:  unknown type
        Name:  600a0b800039c9b500000ab047b457ad
        Asymmetric:  yes
        Current Load Balance:  round-robin
        Logical Unit Group ID:  NA
        Auto Failback:  on
        Auto Probing:  NA

        Paths:
                Initiator Port Name:  10000000c967c830
                Target Port Name:  200400a0b83a8a0c
                Override Path:  NA
                Path State:  OK
                Disabled:  no

                Initiator Port Name:  10000000c967c82f
                Target Port Name:  200500a0b83a8a0c
                Override Path:  NA
                Path State:  OK
                Disabled:  no
```

```
        Target Port Groups:
                ID:  4
                Explicit Failover:  yes
                Access State:  standby
                Target Ports:
                        Name:  200400a0b83a8a0c
                        Relative ID:  0

                ID:  1
                Explicit Failover:  yes
                Access State:  active
                Target Ports:
                        Name:  200500a0b83a8a0c
                        Relative ID:  0
```

Checking the path routing defaults for the logical units

I noticed that with Solaris multi-pathing, the first six LUNs are default active to controller A while the second six LUNs are default active to controller B. The following shell command illustrates this point. We will take this into consideration when assigning drives to pairs when creating the ZFS pool. This approach should obtain the most uniform loading of the FC links and the controllers.

```
# for dev in c4t600A0B80003A8A0B0000096A47B4559Ed0  \
c4t600A0B80003A8A0B0000096E47B456DAd0 \
c4t600A0B80003A8A0B0000096147B451BEd0 \
c4t600A0B80003A8A0B0000096647B453CEd0 \
c4t600A0B80003A8A0B0000097347B457D4d0 \
c4t600A0B800039C9B500000A9C47B4522Dd0 \
c4t600A0B800039C9B500000AA047B4529Bd0 \
c4t600A0B800039C9B500000AA447B4544Fd0 \
c4t600A0B800039C9B500000AA847B45605d0 \
c4t600A0B800039C9B500000AAC47B45739d0 \
c4t600A0B800039C9B500000AB047B457ADd0 \
c4t600A0B800039C9B500000AB447B4595Fd0
do
echo "=== $dev ==="
for> mpathadm show lu /dev/rdsk/$dev | grep 'Access State'
for> done
=== c4t600A0B80003A8A0B0000096A47B4559Ed0 ===
                Access State:  active
                Access State:  standby
=== c4t600A0B80003A8A0B0000096E47B456DAd0 ===
                Access State:  active
                Access State:  standby
=== c4t600A0B80003A8A0B0000096147B451BEd0 ===
                Access State:  active
                Access State:  standby
=== c4t600A0B80003A8A0B0000096647B453CEd0 ===
                Access State:  active
                Access State:  standby
=== c4t600A0B80003A8A0B0000097347B457D4d0 ===
                Access State:  active
                Access State:  standby
=== c4t600A0B800039C9B500000A9C47B4522Dd0 ===
                Access State:  active
                Access State:  standby
=== c4t600A0B800039C9B500000AA047B4529Bd0 ===
                Access State:  standby
```

```
                Access State:  active
=== c4t600A0B800039C9B500000AA447B4544Fd0 ===
                Access State:  standby
                Access State:  active
=== c4t600A0B800039C9B500000AA847B45605d0 ===
                Access State:  standby
                Access State:  active
=== c4t600A0B800039C9B500000AAC47B45739d0 ===
                Access State:  standby
                Access State:  active
=== c4t600A0B800039C9B500000AB047B457ADd0 ===
                Access State:  standby
                Access State:  active
=== c4t600A0B800039C9B500000AB447B4595Fd0 ===
                Access State:  standby
                Access State:  active
```

Creating the pool

Creating the pool took less than three seconds and resulted in 1.6TB of usable space.

```
# zpool create Sun_2540 \
  mirror \
    c4t600A0B80003A8A0B0000096A47B4559Ed0 \
    c4t600A0B800039C9B500000AA047B4529Bd0 \
  mirror \
    c4t600A0B80003A8A0B0000096E47B456DAd0 \
    c4t600A0B800039C9B500000AA447B4544Fd0 \
  mirror \
    c4t600A0B80003A8A0B0000096147B451BEd0 \
    c4t600A0B800039C9B500000AA847B45605d0 \
  mirror \
    c4t600A0B80003A8A0B0000096647B453CEd0 \
    c4t600A0B800039C9B500000AAC47B45739d0 \
  mirror \
    c4t600A0B80003A8A0B0000097347B457D4d0 \
    c4t600A0B800039C9B500000AB047B457ADd0 \
  mirror \
    c4t600A0B800039C9B500000A9C47B4522Dd0 \
    c4t600A0B800039C9B500000AB447B4595Fd0
```

Status of the pool

The pool has yet to detect any data errors.

```
# zpool status
  pool: Sun_2540
 state: ONLINE
 scrub: scrub completed with 0 errors on Mon Feb 25 13:54:54 2008
config:

        NAME                                        STATE     READ WRITE CKSUM
        Sun_2540                                    ONLINE       0     0     0
          mirror                                    ONLINE       0     0     0
            c4t600A0B80003A8A0B0000096A47B4559Ed0   ONLINE       0     0     0
            c4t600A0B800039C9B500000AA047B4529Bd0   ONLINE       0     0     0
          mirror                                    ONLINE       0     0     0
            c4t600A0B80003A8A0B0000096E47B456DAd0   ONLINE       0     0     0
            c4t600A0B800039C9B500000AA447B4544Fd0   ONLINE       0     0     0
          mirror                                    ONLINE       0     0     0
```

```
            c4t600A0B80003A8A0B0000096147B451BEd0  ONLINE       0     0     0
            c4t600A0B800039C9B500000AA847B45605d0  ONLINE       0     0     0
          mirror                                   ONLINE       0     0     0
            c4t600A0B80003A8A0B0000096647B453CEd0  ONLINE       0     0     0
            c4t600A0B800039C9B500000AAC47B45739d0  ONLINE       0     0     0
          mirror                                   ONLINE       0     0     0
            c4t600A0B80003A8A0B0000097347B457D4d0  ONLINE       0     0     0
            c4t600A0B800039C9B500000AB047B457ADd0  ONLINE       0     0     0
          mirror                                   ONLINE       0     0     0
            c4t600A0B800039C9B500000A9C47B4522Dd0  ONLINE       0     0     0
            c4t600A0B800039C9B500000AB447B4595Fd0  ONLINE       0     0     0

errors: No known data errors
```

Iozone results prior to StorageTek 2540 Tuning:

These are the results before the cache mirroring was disabled, and the cache flushing was disabled. Notice that a 64GB test file was used to eliminate most effects from ARC caching.

```
      Iozone: Performance Test of File I/O
              Version $Revision: 3.283 $
            Compiled for 64 bit mode.
            Build: Solaris10gcc-64

      Contributors:William Norcott, Don Capps, Isom Crawford, Kirby Collins
                   Al Slater, Scott Rhine, Mike Wisner, Ken Goss
                   Steve Landherr, Brad Smith, Mark Kelly, Dr. Alain CYR,
                   Randy Dunlap, Mark Montague, Dan Million,
                   Jean-Marc Zucconi, Jeff Blomberg, Benny Halevy,
                   Erik Habbinga, Kris Strecker, Walter Wong.

      Run began: Thu Feb 14 16:35:51 2008

      Auto Mode
      Using Minimum Record Size 64 KB
      Using Maximum Record Size 512 KB
      Using minimum file size of 33554432 kilobytes.
      Using maximum file size of 67108864 kilobytes.
      Command line used: iozone -a -i 0 -i 1 -y 64 -q 512 -n 32G -g 64G
      Output is in Kbytes/sec
      Time Resolution = 0.000001 seconds.
      Processor cache size set to 1024 Kbytes.
      Processor cache line size set to 32 bytes.
      File stride size set to 17 * record size.

              KB  reclen   write rewrite    read    reread
        33554432      64  150370  113779   454731   456158
        33554432     128  147032  181308   455496   456239
        33554432     256  148182  169944   454192   456252
        33554432     512  153843  194189   473982   516130
        67108864      64  151047  111227   463406   456302
        67108864     128  148597  159236   456959   488100
        67108864     256  148995  165041   463519   453896
        67108864     512  154556  166802   458304   456833
```

Iozone results after StorageTek 2540 Tuning:

These are the final results. Not much to complain about here!

```
      Iozone: Performance Test of File I/O
```

```
        Version $Revision: 3.283 $
      Compiled for 64 bit mode.
      Build: Solaris10gcc-64

Contributors:William Norcott, Don Capps, Isom Crawford, Kirby Collins
             Al Slater, Scott Rhine, Mike Wisner, Ken Goss
             Steve Landherr, Brad Smith, Mark Kelly, Dr. Alain CYR,
             Randy Dunlap, Mark Montague, Dan Million,
             Jean-Marc Zucconi, Jeff Blomberg, Benny Halevy,
             Erik Habbinga, Kris Strecker, Walter Wong.

Run began: Sat Feb 16 11:19:42 2008

Auto Mode
Using Minimum Record Size 64 KB
Using Maximum Record Size 512 KB
Using minimum file size of 33554432 kilobytes.
Using maximum file size of 67108864 kilobytes.
Command line used: iozone -a -i 0 -i 1 -y 64 -q 512 -n 32G -g 64G
Output is in Kbytes/sec
Time Resolution = 0.000001 seconds.
Processor cache size set to 1024 Kbytes.
Processor cache line size set to 32 bytes.
File stride size set to 17 * record size.

        KB  reclen   write rewrite    read    reread
  33554432      64  279863  167138   458807   449817
  33554432     128  265099  250903   455623   460668
  33554432     256  265616  259599   451944   448061
  33554432     512  278530  294589   522930   471253
  67108864      64  273739  168477   455085   455951
  67108864     128  263852  289383   455225   456217
  67108864     256  260685  289844   452569   450746
  67108864     512  273984  307212   484305   453194
```